Capstone Project Week1

Introduction

This project is going to use Jupyter Notebook to analyze GEO data of a Sydney suburb named Beecroft.

As part of the Data Science assignment, it'll use public available data as well as various libraries and tools. The Foursquare location data will be used during the exercise.

Business Problem

Often people feel lost when trying to make decision based on location. The data is either not available or not well organized. This project is trying to resolve this problem by analysing various related data.

Based on the detail analysis, interesting party may make business decision with more information and confidence. For example, a potential property investor may purchase a property in suburbs which have more desired facilities. A small business may open a coffee shop in a suburb to maximize the profit.

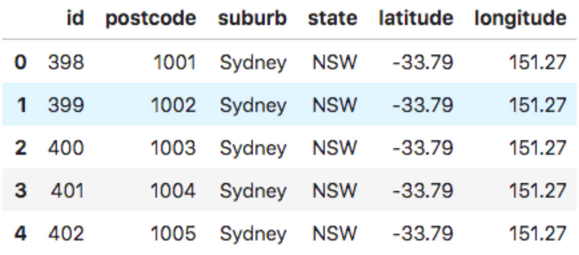
The target audience for the report are home buyers and business investors.

Data Description

The Sydney metro data used is from this link:

https://www.prospectshop.com.au/Files/SydneyMetro_Postcodes.xls

The sample data:



	id	postcode	suburb	state	latitude	longitude
0	398	1001	Sydney	NSW	-33.79	151.27
1	399	1002	Sydney	NSW	-33.79	151.27
2	400	1003	Sydney	NSW	-33.79	151.27
3	401	1004	Sydney	NSW	-33.79	151.27
4	402	1005	Sydney	NSW	-33.79	151.27

The original data format will be transformed to work together with the other APIs. The Data filtering phase:

- Format the column headers
- Remove duplicate suburbs since one suburb may have more entries of postcodes due to the post box.

- Remove suburb without GEO data
- Remove PO box entries which have postcode start with 1000

The location and venue data come from the Foursquare.

Postcode data:

https://gist.github.com/randomecho/5020859

Need to import the data from the above URL into MySQL and then export the data as .csv file. This involves to install MySQL in a Ubuntu VM, execute the SQL statement from the above link, export data as .csv.